Maths Test13
Observations and Comments

Larry R Nelson[1]

21 March 2022

The 13 items

With one exception, the items in this quiz appear fairly straightforward to me. The exception would be the second question, an item that requires knowledge of the number of days in two months of the year. This question seems to me to be out of line with the others, that is, dissimilar, an "oddball". Of course, I don't know the instructional syllabus underlying the test, but, to me, maths instruction would not ordinarily involve a question such as this one.

Missing data

In this dataset, a response code of 9 was used whenever a student did not answer a test item, that is, when a response was totally missing.

In Lertap5, missing responses will generally be tabulated in a Lertap5 "**other**" column or row, depending on the worksheet report involved.

The results in Figure 1 were obtained when I was looking at the original Excel workbook having response records for 2,976 students.

Options->	1	2	3	4	other
Q1	6%	10%	75%	5%	4%
Q2	35%	21%	18%	12%	13%
Q3	9%	11%	58%	15%	6%
Q4	50%	13%	16%	13%	8%
Q5	7%	64%	20%	5%	5%

Figure 1

The table seen in Figure 1 is an excerpt from a Lertap5 "Stats1b" worksheet report. In this example, the "other" figures are seen in the last column.

The table indicates that 13% of the students did not answer Q2.

I found that several questions were missing 10% or more responses: Q2, Q7, Q9, Q11, Q12, and Q13 – the latter item, Q13, had the most missing data (15%).

I also found that 34 students did not answer any of the items.

[1] Lertap5@gmail.com or l.nelson@curtin.edu.au

After looking further into missing data, I decided that an analysis of test results would be optimal if I used results only for those students who had answered at least eight of the thirteen test items. This left me with a new Excel workbook having item responses from 2,727 students.

Item difficulty and Stats1b reports

In classical test theory (CTT), the "difficulty" measure for an item is equal to the percentage (or proportion) of students able to answer the item correctly.

In a Lertap5 Stats1b report, such as that seen in Figure 2 below, the correct answer to each item has its percentage figure underlined. This is the item's classical index of difficulty, repeated as a proportion in the Difficulty column (78%, for example, is 0.78 as a proportion).

Difficult items are those with the *lowest* percentage-correct figure; easy items have the *highest* percentage correct.

In Figures 1 and 2, Q1 was the easiest item (difficulty 0.78); Q2 was the hardest (difficulty 0.38). This terminology is the reverse of what might seem more logical – we'd ordinarily expect difficult items to have a difficulty index higher than that for easy items. Some have tried to correct this historic anomaly by changing "difficulty index" to "facility index", but, by and large, this idea has not taken hold.

Item discrimination

The item discrimination values in a Stats1b report, as seen in Figure 2, are correlation coefficients. The discrimination for an item is computed by having Lertap5 create two scores for each student, and then correlating them.

One of the scores will be zero or one – one if the student got the item correct, zero otherwise. The other score will (almost always) simply be the number of other items which the student correctly answered.

For example, suppose Student 1 answered Q1 correctly, and, over the remaining 12 items, got six of them correct. The two scores for this student would be (1,6).

This is done for every student. The correlation between the two scores is then taken as the item discrimination index.

An item's discrimination value will be high, close to the maximum of 1.00, when those students getting an item correct have a high score on the other test items.

Figure 3 shows that there was a cluster of items having discrimination values close to 0.40 and above. Q2 and Q13 had low item discrimination.

A common objective in testing is to have items with good discrimination, above 0.30 for example. Such tests will have high reliability. More about item discrimination is found in Chapter7 of the Lertap5 manual.

Options->	1	2	3	4	other	Difficulty	Discrimination	?
Q1	6%	9%	78%	5%	2%	0.78	0.26	
Q2	37%	22%	19%	13%	10%	0.37	0.13	
Q3	10%	10%	62%	15%	3%	0.62	0.43	
Q4	53%	12%	16%	14%	4%	0.53	0.52	
Q5	6%	68%	20%	5%	2%	0.68	0.50	
Q6	48%	15%	14%	19%	4%	0.48	0.49	
Q7	9%	12%	60%	13%	7%	0.60	0.38	
Q8	7%	51%	11%	29%	2%	0.51	0.43	
Q9	16%	45%	21%	16%	3%	0.45	0.41	
Q10	16%	13%	58%	10%	3%	0.58	0.43	
Q11	14%	21%	50%	11%	4%	0.50	0.40	
Q12	9%	43%	21%	22%	5%	0.43	0.44	
Q13	10%	24%	18%	41%	8%	0.41	0.17	

... Stats1f **Stats1b** Stats1ul Stats2f Stats2b Stats

Figure 2

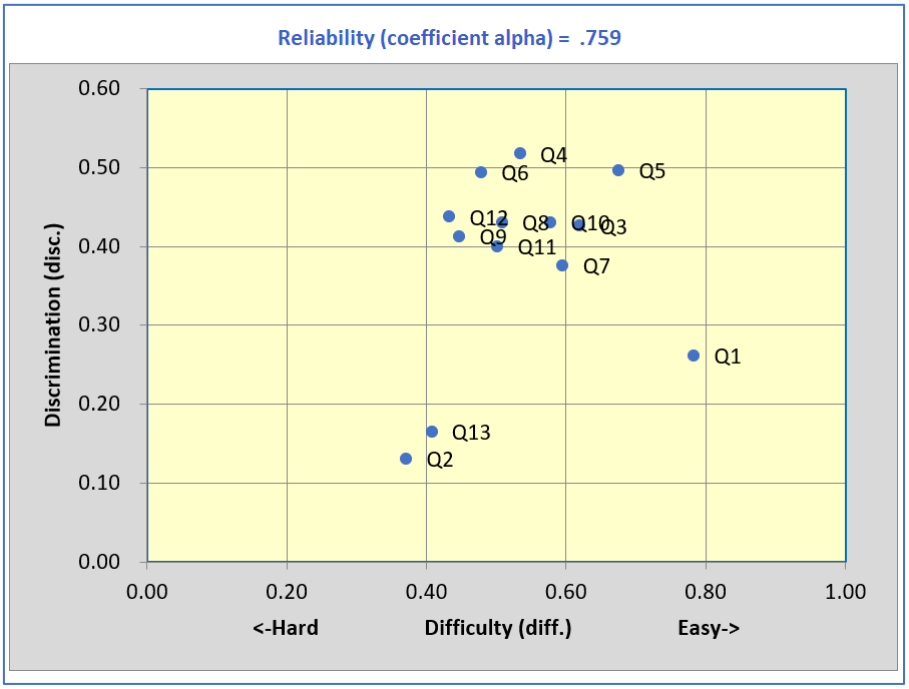

Figure 3

Discussing results (the actual items may be seen here)

Often instructors will place easy items at the beginning of a quiz or test, thinking, perhaps, that doing so may help students relax.

I myself would say that Q1, asking students to sum two four-digit numbers (without being able to use a calculator), might indeed be an easy item. Was it so for these students?

Well, almost 80% did get this item right. That may sound good, but I think, were I the instructor, I myself might have hoped for an even better result.

But then look at Q2, the item having to do with the number of days in certain months of the year. So much for putting easy items towards the beginning of a quiz, eh? Q2 was very hard for the students. Q1 was much easier.

Now, looking at the items, we might think that the student abilities required to answer some sets of consecutive items would be the same. For example …

Q3 and Q4 are highly similar; Q5 and Q6 are similar; Q7, Q8, and Q9 seem to me to be tapping the same ability; Q12 and Q13, two division items, would require the same ability, wouldn't you think?

But look at the results. Consider Q5 (68% correct) and Q6 (48%), quite a difference. Q7, Q8, and Q9 go 60% correct, 51%, then 45%, again quite a difference over basically similar items. Q12 and Q13 (on the other hand) do have similar results.

Getting a total test score

My discussion to this point has been at "the item level", that is, focused on outcomes item by item.

There are certainly many times when instructors will interpret results based exclusively on the item statistics alone. In this case, for example, assuming that students are expected to know the number of days in the months of June and July, and to use such knowledge to answer Q2, then, given the results seen above, the instructor might want to spend more class time on this topic as the students did not do well on Q2.

Of course, there are also times when instructors, and especially administrators, may want a single statistic that can serve to summarise the overall outcome resulting from the administration of a quiz or test.

The total test score is called on for this purpose.

Lertap5, and other item analysis programs, routinely calculate total test scores, one for each student. The usual "method" is to award one point for each correct answer, and then sum the points to get the total. This is "number right" scoring.

For our 13-item test, total scores would range from 0 (zero) to 13. Results from Lertap5 are seen in Figure 4.

n	2,727
Min	0.00
Median	7.00
Mean	6.94
Max	13.00
s.d.	3.21
var.	10.29
Range	13.00
IQRange	6.00
Skewness	0.06
Kurtosis	-1.06
MinPos	0.00
MaxPos	13.00

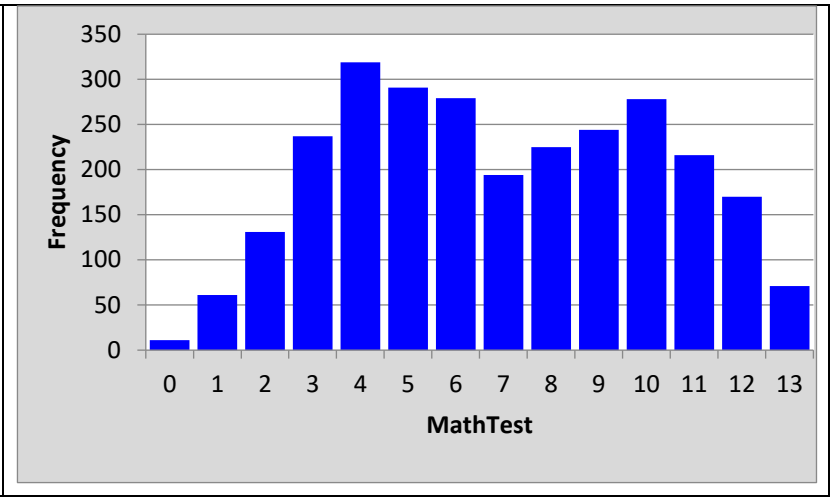

Figure 4

Regional differences?

The students who took this test were located in four distinct regional administrative zones.

The following figures breakout the total score by zones (Figs. 5 and 6), and then present a response summary, by zones, for the first question, Q1 (Fig. 7).

MathTest	Zone1	Zone2	Zone3	Zone4
n	699	688	556	784
Min	0.00	0.00	0.00	0.00
Median	6.00	7.00	5.00	8.00
Mean	6.62	7.36	5.38	7.95
Max	13.00	13.00	13.00	13.00
s.d.	3.23	3.08	2.87	3.06
var.	10.41	9.46	8.26	9.37
Range	13.00	13.00	13.00	13.00
IQRange	5.00	5.00	4.00	4.00
Skewness	0.32	-0.12	0.54	-0.32
Kurtosis	-0.99	-1.04	-0.44	-0.82
MinPos	0.00	0.00	0.00	0.00
MaxPos	13.00	13.00	13.00	13.00

Figure 5

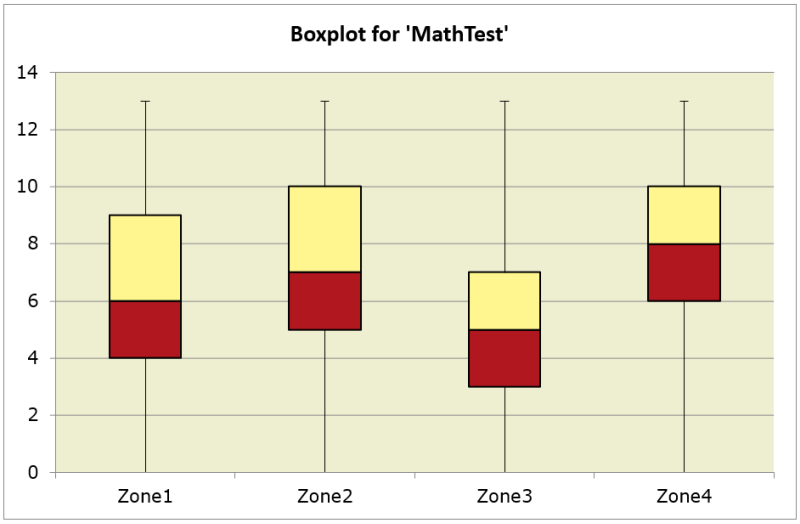

Figure 6[2]

Q1	1	2	3	4	Other	n	mean	s.d.
Zone1	5%	7%	**80%**	6%	2%	699	0.80	0.40
Zone2	3%	7%	**83%**	5%	2%	688	0.83	0.37
Zone3	13%	15%	**63%**	7%	2%	556	0.63	0.48
Zone4	4%	10%	**84%**	2%	1%	784	0.84	0.37
					F, sig, eta²:	*36.64*	*0.00*	*0.04*

Figure 7

Zone 3's results were the weakest, with Figure 7 above highlighting what could be of concern to teachers and administrators in this zone – the 556 students in Zone3 were considerably less successful when it came to solving Q1, the easiest item on the test, simple addition.

Test reliability

Imagine this common situation: a country's teachers are gathered together in a regional workshop, or perhaps a national conference. A session presenter comes forward, introduces Test13, explains the rationale behind the test items, goes on to present test results (much as I have above), and discusses possible implications for maths education and assessment.

At question time, someone asks: *What was the test's reliability?*

Reliability has to do with replicability. If the test were used again with the same students, would results be the same?

It's an important question indeed. It's so important that the presenter might apologize for overlooking it and take the audience back to Figure 3 where the value of coefficient alpha is given (0.759).

Interpreting common reliability estimates, such as alpha, relies on a knowledge of correlation coefficients. A high alpha value means that, were we able to give

[2] Lertap5 boxplots are discussed here.

students a chance to sit the test again, their scores from the two testings would be similar.

Correlation coefficients range in value from -1.00 to +1.00. In the unlikely event that the two test scores came out to be identical for each of the students, the test would be "perfectly reliable" as it gave the same scores each time.

If the correlation came out to be 0.00, test scores would be perfectly unreliable, useless for almost every purpose we might think of.

For more information about coefficient alpha, Chapter 7 of the Lertap5 manual is a suggested reference.

At times test developers will develop "parallel forms" of a test, such as "FormA" and "FormB", give both forms to the same group of students, and correlate student scores from both forms, resulting in a "parallel forms reliability estimate". This link leads to an authentic example.

It might be easy to take Test13 and make a parallel form. We'd get in and change the numeric values seen in each item. Q1 might, for example, ask for the sum of 2015 and 6106. We'd make similar changes to each of the items, creating another 13-item test, a "parallel form". Of course, it would be best if the developers of Test13 did this as the rationale behind each of the present items might not be as simple as first appearances suggest to us.

Wu et al. (2016) present more extensive discussions of test reliability.

Test13's reliability

Look above, if you would, at Figure 3. Lertap5 has found the reliability of Test13, when given to the 2,727 students, to be 0.759, call it 0.76.

This isn't too bad at all for a short test. But it could be better.

Q2 and Q13 have been found to have poor discrimination, they stick out in the figure, their discrimination values are below 0.20. For some reason they're not assisting us, meaning that they didn't perform as did the other items when it comes to identifying the best students. An item is said to "discriminate" when it helps pick out the strongest students, those who appear to have the greatest knowledge of the subject covered by the test. Q2 and Q13 did not discriminate; they did nothing, or very little at the most, in helping to identify the best students.

Now look at Figure 8. Q2 and Q13 are gone.

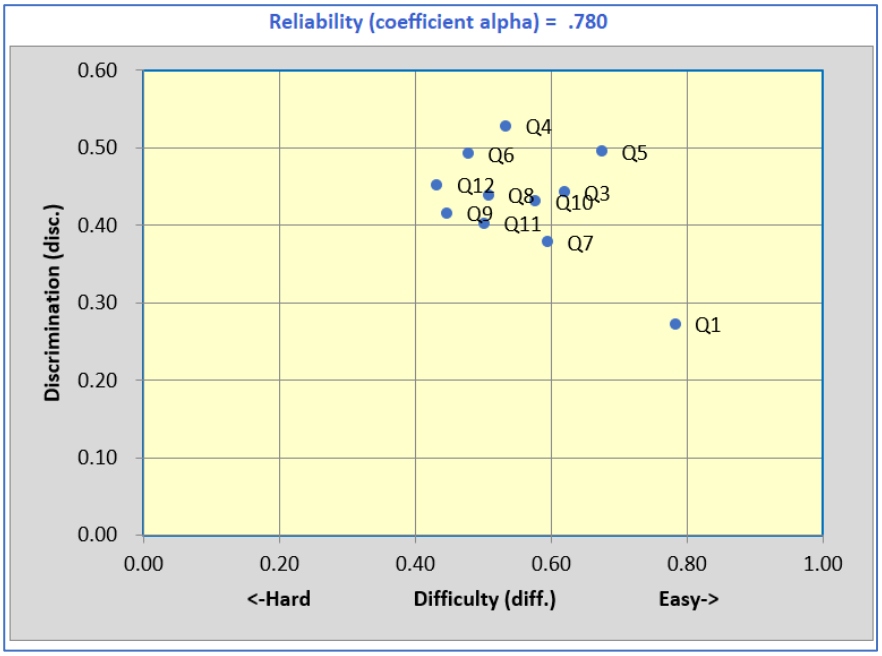

Figure 8

With Q2 and Q13 taken out of the test, Figure 8 shows that test reliability increased to 0.78, a very modest but nonetheless useful boost over the 0.76 seen in Figure 3 when Q2 and Q13 were included.

In the business of test development and application, an 11-item multiple-choice test with a reliably just shy of 0.80 would probably not be criticized[3].

As an aside, I can see why the removal of Q2 would improve reliability. It requires knowing how many days there are in the months of June and July. I ask myself why such a question would be included on a maths test? The results I've found from Lertap5 show that Q2 has served to lower the test's reliability.

A mystery, however, regards Q13. It's Q12 with different numbers, and Q12 had good discrimination. Why has Q13's discrimination turned out to be so low? I certainly haven't yet figured out the problem with this item, but it is clear that test reliability improves when it, along with Q2, is omitted from the calculation of the total test score.

References

The Lertap5 manual, Chapter 7. *A top read*.

Figure 5 was made using this Lertap5 option. Figure 7 made use of this option.

Wu, M, Hak Ping Tam, & Tsung-Hau Jen (2016), *Educational Measurement for Applied Researchers*: *Theory into Practice*. Springer Singapore.

[3] In other words, 0.78 reliability would often be accepted as good enough, especially as this test is so short.